DD2434 – Advanced Machine Learning

Lecture 2: **Probabilistic approach – fundamentals**

Pawel Herman

Computational Science and Technology (CST)

KTH Royal Institute of Technology

Course content

- Aims and scope of this part of the course

I. Probabilistic regression

- *Lecture 2:* Probabilistic approach (objects and inference)
- *Lecture 3:* Linear regression
- *Lecture 4:* Kernels and introduction to Gaussian processes
- *Lecture 5:* Gaussian processes

II. Probabilistic representation learning

- *Lecture 6:* Representation learning

Course content

- Learning activities
  - 5 lectures
  - 2 exercise sessions (fun with Gaussians ;-))
  - 1 assignment
  - project based on a research paper

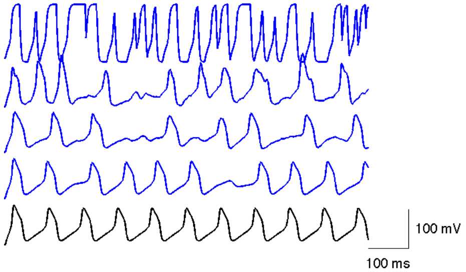
Short outline for today

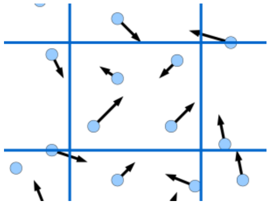
1. Introduction – why is probability theory relevant in ML?
2. Probabilistic approach to ML – principles (*recap*).
3. Probability basics (tbc).

Introduction

Probabilistic approach

- Ubiquitous nature of uncertainty
  - imprecision, noise in data,
  - errors , missing information/data
  - gaps in knowledge, simplified description



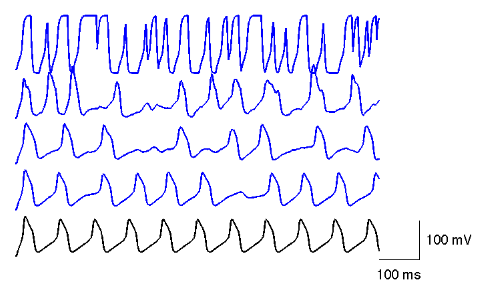


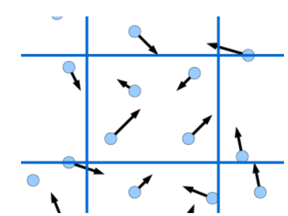
Introduction

Probabilistic approach

➢ Ubiquitous nature of uncertainty

- imprecision, noise in data,
- errors , missing information/data
- gaps in knowledge, simplified description



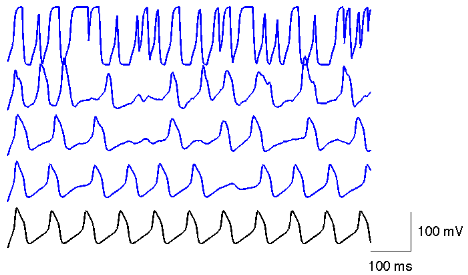


"The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account." (Laplace)

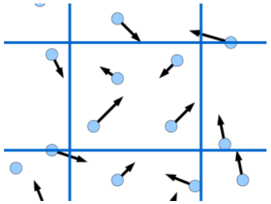
Introduction

Probabilistic approach

- Ubiquitous nature of uncertainty
  - imprecision, noise in data
  - errors , missing information/data
  - gaps in knowledge, simplified description



- Probability theory provides a framework for modelling, reasoning etc. under uncertainty
  - unified, universal, intuitive, interpretable
  - beyond randomness, it is about uncertainty!
  - p. distributions as carriers or information (Jaynes, 2003)



Probabilistic perspective in ML

- Statistical ML: constructing stochastic models
  - fully probabilistic description and inference
  - theoretical assumptions, mathematical tractability, rigour
  - parameters estimated from observed data (learning)
  - interpretability and extra insights
  - machinery to propagate and account for uncertainty effects

Probabilistic perspective in ML

- Statistical ML: constructing stochastic models
  - fully probabilistic description and inference
  - theoretical assumptions, mathematical tractability, rigour
  - parameters estimated from observed data (learning)
  - interpretability and extra insights
  - machinery to propagate and account for uncertainty effects

BUT: *can be very hard for large-scale problems* and

difficult to derive solutions in a closed form

Probabilistic perspective in ML

- Statistical ML: constructing stochastic models
  - fully probabilistic description and inference
  - t … igour
  - p
  - i
  - n … effects

> “ *(statistical ML) provides the basis for learning algorithms that directly manipulate probabilities, as well as a framework for analyzing the operation of other algorithms that do not explicitly manipulate probabilities.*” Mitchell

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
    - product and sum probability rules, Bayesian theorem

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
    - product and sum probability rules, Bayesian theorem
    - the power of marginalisation

$$p(x_1, x_2 .., x_{n-1}) = \int p(x_1, x_2, .., x_n) dx_n$$

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
  - Combine uncertain knowledge with data to reduce uncertainty (based on evidence from observations)
    - batch or sequence where posterior is iteratively updated

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
  - Combine uncertain knowledge with data to reduce uncertainty (based on evidence from observations)
  - Two levels of inference:

    parameter estimation and model selection (see Lecture 3)

Other ML approaches

- Probabilistic narration is not the only one in ML

concept learning
based on symbolic
representations
(inductive learning
hypothesis)

Other ML approaches

- Probabilistic narration is not the only one in ML

concept learning (inductive learning hypothesis)

rule learning (e.g., Inductive Logic Programming)

and rule-based learning

instance based learning (k-means)

Other ML approaches

- Probabilistic narration is not the only one in ML

concept learning (inductive learning hypothesis)

instance based learning (k-means)

rule learning (e.g., Inductive Logic Programming) and rule-based learning

computational learning theory (PAC or Mistake-Bound learning model)

Other ML approaches

- Probabilistic narration is not the only one in ML

concept learning (inductive learning hypothesis)

instance based learning (k-means)

rule learning (e.g., Inductive Logic Programming) and rule-based learning

computational learning theory (PAC or Mistake-Bound learning model)

decision tree learning

reinforcement learning

Other ML approaches

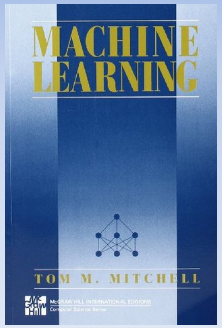
- Probabilistic narration is not the only one in ML

concept lea
based on sy
representa
(inductive le
hypothes

You also may have heard about

- artificial neural networks, deep architectures
- search heuristics, e.g. genetic algorithms
- support vector machines

  (origins in comp learning theory)
- etc.



nforcement
learning

instance based
learning (k-means)

and rule-based
learning

(PAC or Mistake-Bound
learning model)

sion tree
learning

Other ML approaches

- Probabilistic narration is not the only one in ML

concept lea
based on sy
representa
(inductive le
hypothes

nforcement
learning

You also may have heard about

- artificial neural networks, deep architectures
- search heuristics, e.g. genetic algorithms
- support vector machines (origins in comp learning theory)
- etc.

ML comes in many flavours

instance based
learning (k-means)

and rule-based
learning

(PAC or Mistake-Bound
learning model)

sion tree
learning

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

I. The ratio of outcomes in repeated trials.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random realisation.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

III. There does not have to be an experiment for declaring probability.

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random realisation.

III. Parameters can only be deduced from data (likely outcome of exp.)

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

III. There does not have to be an experiment for declaring probability.

IV. Can incorporate prior knowledge, probabilities can be updated.

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random realisation.

III. Parameters can only be deduced from observed data (likely outcome of exp.)

IV. Each repeated experiment starts from ignorance.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)
II. The data is fixed, models have probabilities
III. There does not have to be an experiment for declaring probability.
IV. Can incorporate prior knowledge, probabilities can be updated.
V. Estimators are good for available data.

I. The ratio of outcomes in repeated trials.
II. There is a true model and the data is a random realisation.
III. Parameters can only be deduced from observed data (likely outcome of exp.)
IV. Each repeated experiment starts from ignorance.
V. Estimators are averaged across many trials.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)
II. The data is fixed, models have probabilities
III. There does not have to be an experiment for declaring probability.
IV. Can incorporate prior knowledge, probabilities can be updated
V. Estimators are good for available data.
VI. Probability of a hypothesis given the data (posterior distribution).

I. The ratio of outcomes in repeated trials.
II. There is a true model and the data is a random realisation.
III. Parameters can only be deduced from observed data (likely outcome of exp.)
IV. Each repeated experiment starts from ignorance.
V. Estimators are averaged across many trials.
VI. Probability of the data given hypothesis (likelihood, sampling dist.).

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)
II. The data is fixed, models have probabilities
III. There does not have to be an experiment for declaring probability.
IV. Can incorporate prior knowledge, probabilities can be updated
V. Estimators are good for available data.
VI. Probability of a hypothesis given the data.

VII. All variables/parameters have distribution.

I. The ratio of outcomes in repeated trials.
II. There is a true model and the data is a random realisation.
III. Parameters can only be deduced from observed data (likely outcome of exp.)
IV. Each repeated experiment starts from ignorance.
V. Estimators are averaged across many trials.
VI. Probability of the data given hypothesis.

VII. Parameters are fixed unknowns that can be point estimated from repeated trials.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

III. There does not
declaring proba

IV. Can incorporate
probabilities can be updated

V. Estimators are good for available data.

VI. Probability of a hypothesis given the data.

VII. All variables/parameters have distribution.

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random

uced from
ne of exp.)

IV. Each repeated experiment starts from ignorance.

V. Estimators are averaged across many trials.

VI. Probability of the data given hypothesis.

VII. Parameters are fixed unknowns that can be point estimated from repeated trials.

Why isn't everyone Bayesian? *(Efron, 1986)*

Learning as inference – recap (*c.f.* DD2431, L6 G. Salvi)

- Learning distributions
  - *"integrating data with domain knowledge of model environment"*

$$y = f(x) : X \to Y$$

$$\{(x_1, t_1), \ldots, (x_N, t_N)\}$$

Learning implies the parameter identification of the model adopted to account for the mapping $X \to Y$

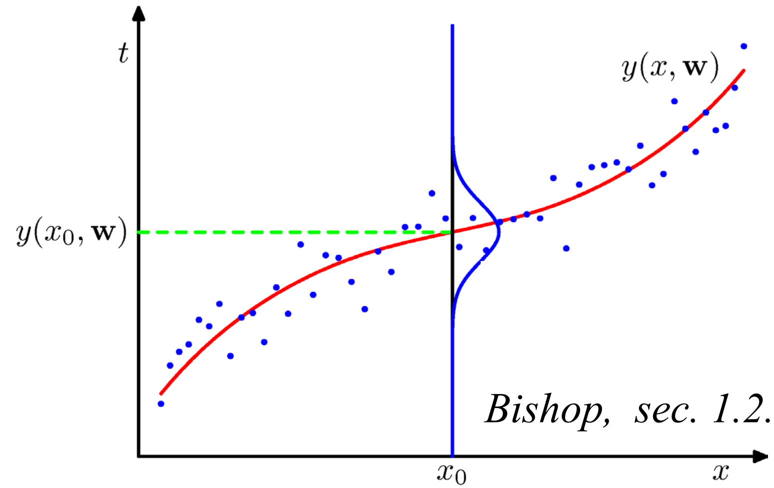
Inference is about using the model to predict the output corresponding to the sample input: *what is the model's response for* $x = x^*$?

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x, \mathbf{w}) : X \to Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$



Bishop, sec. 1.2.5

Learning as inference – recap

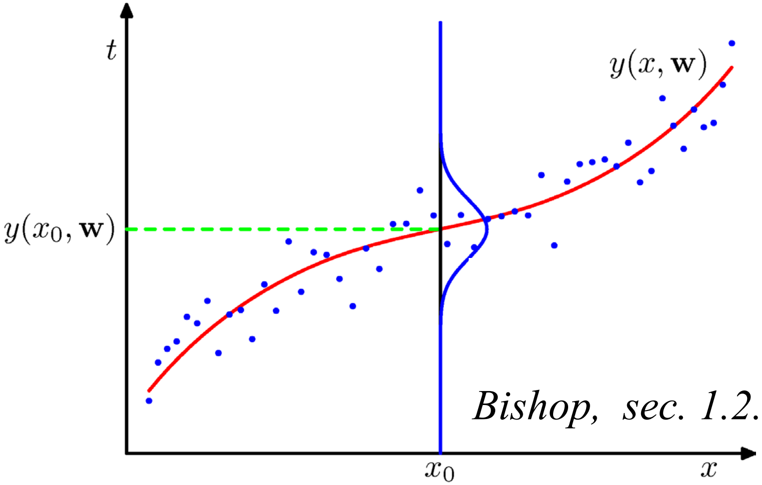
- Learning distributions
  - curve-fitting example: $y(x, \mathbf{w}) : X \to Y$ (let's assume polynomial)

$$t = y(x, \mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

<u>Remarks about notation:</u>

1) here we deal with one-dim input, x and output, t
2) parameters $\mathbf{w}$ still constitute a vector (e.g. could be polynomial coefficients)
3) $\mathbf{x}$ and $\mathbf{t}$, refer to the collection of all inputs and outputs, conceptually corresponding to D_x and D_t, but in the vector form, so $\mathbf{x} \to \mathbf{t}$.



Bishop, sec. 1.2.5

Learning as inference – recap

- Learning distributions

  - curve-fitting example: $y(x,\mathbf{w}) : X \rightarrow Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1 , \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1 , \ldots, t_N)^{\mathrm{T}}$

  - probabilistic framework

$p(\mathbf{w})$ – a prior probability distribution

$p(\mathcal{D} \mid \mathbf{w})$ – the likelihood function for $\mathcal{D} = \mathbf{t}$
(not a probability distrib. over $\mathbf{w}$, just a conditional probability)

$p(\mathbf{w} \mid \mathcal{D})$ – a posterior probability

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \to Y$ (let's assume polynomial)

observatio $_N)^{\mathrm{T}}$

  - probab

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

$p(\mathbf{w}$

$p(\mathcal{D} \mid \mathbf{w})$ – the likelihood function for $\mathcal{D} = \mathbf{t}$
(not a probability distrib. over $\mathbf{w}$, just a conditional probability)

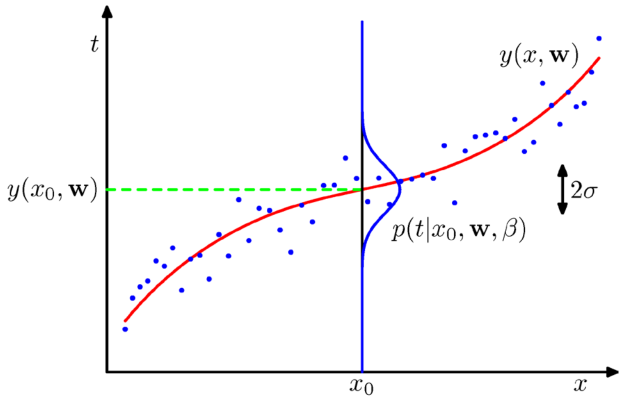
$p(\mathbf{w} \mid \mathcal{D})$ – a posterior probability

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \rightarrow Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i) : i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^T$, $\mathbf{t} = (t_1, \ldots, t_N)^T$



Bishop, sec. 1.2.5

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x, \mathbf{w}) : X \rightarrow Y$ (let's assume polynomial)

$$t = y(x, \mathbf{w}) + \varepsilon$$

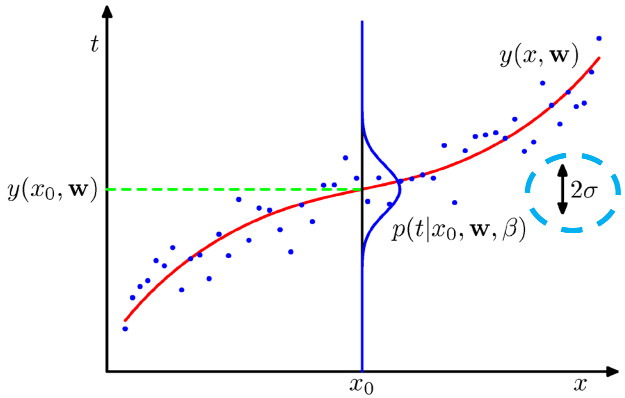
observations $\{(x_i, t_i) : i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

Uncertainty (noise) in target data:

$$p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$$

↑

precision



Bishop, sec. 1.2.5

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \rightarrow Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

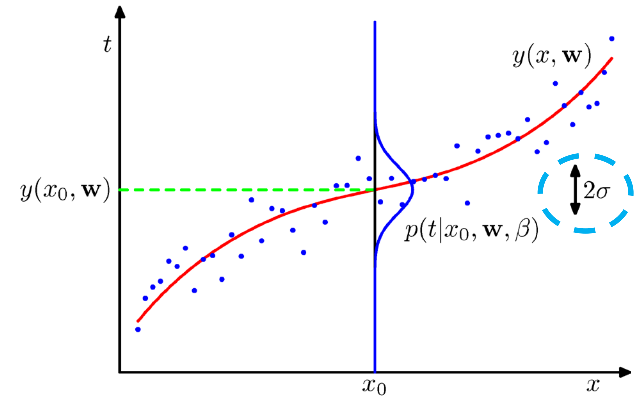
observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1 , ..., x_N)^T$, $\mathbf{t} = (t_1 , ..., t_N)^T$

Uncertainty (noise) in target data:

$$\underbrace{p(t \mid x, \mathbf{w}, \beta)}_{\text{predictive distribution}} = \mathcal{N}(t \mid y(x,\mathbf{w}), \beta^{-1})$$

predictive **distribution**
(not a point estimate)

precision



Bishop, sec. 1.2.5

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \to Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1, \dots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \dots, t_N)^{\mathrm{T}}$

SOLUTION 1

We want to find parameters to be able to use predictive distribution, $p(t|x)$, and infer the target:

$$\mathrm{E}[t \mid x] = \int t\, p(t \mid x, \mathbf{w}_{\mathrm{opt}}, \beta_{\mathrm{opt}})\,\mathrm{d}t$$

Learning as inference – recap

So, we follow the max likelihood approach

ML function for t (i.i.d.) under Gaussian noise $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y_n(x_n, \mathbf{w}), \beta^{-1})$$

Learning as inference – recap

So, we follow the max likelihood approach

ML function for t (i.i.d.) under Gaussian noise $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y_n(x_n, \mathbf{w}), \beta^{-1})$$

The log-likelihood:

$$\boxed{\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)}$$

Learning as inference – recap

So, we follow the max likelihood approach

ML function for t (i.i.d.) under Gaussian noise $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y_n(x_n, \mathbf{w}), \beta^{-1})$$

The log-likelihood:

$$\boxed{\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)}$$

Still, input x is one-dim and $\mathbf{x}$ & $\mathbf{t}$ refer to the collection of all data points.

Learning as inference – recap

Maximum likelihood (ML) estimate:

Maximise $$\ln p(\mathbf{t}\,|\,\mathbf{x},\mathbf{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

Learning as inference – recap

Maximum likelihood (ML) estimate:

Maximise $\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \left\{ \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right\}$$

the sum-of-squares error function

under the assumption of Gaussian noise : $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

Learning as inference – recap

Maximum likelihood (ML) estimate:

Maximise $$\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \left\{ \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 \right\}$$

$$\beta_{\mathrm{ML}}^{-1} = \frac{1}{N} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n \}^2$$

Learning as inference – recap

Maximum likelihood (ML) estimate:

Maximise $$\ln p(\mathbf{t}\,|\,\mathbf{x},\mathbf{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$$

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}}\left\{\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2\right\}$$

$$\beta_{\mathrm{ML}}^{-1} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}_{\mathrm{ML}})-t_n\}^2$$

$$\boxed{t_{out} = \mathrm{E}[t\,|\,x_{in}] = \int t\,p(t\,|\,x_{in},\mathbf{w}_{\mathrm{ML}},\beta_{\mathrm{ML}})\,\mathrm{d}t}$$

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}): X \rightarrow Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

Solution 2

Parameters of the model could also be random variables with a prior distribution, $p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$

Now, we must **maximise the posterior** $p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$, i.e. find the most probable $\mathbf{w}$ given data:

$$\max p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta)\ p(\mathbf{w} \mid \alpha)$$

Learning as inference – recap

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \to Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

Solution 2

Parameters of the model could also be random variables with a prior distribution, $p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$

Now, we must **maximise the posterior** $p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$, i.e. find the most probable $\mathbf{w}$ given data:

$$\max p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta)\ p(\mathbf{w} \mid \alpha)$$

Again, $\mathbf{x}$ & $\mathbf{t}$ refer to the data collection, not individual input or outputs

Learning as inference – recap

Maximum posterior (MAP) estimate:

Maximise $\ln p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$

$$\downarrow$$

$$\boxed{\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \left\{ \frac{\beta}{2} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 + \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \right\}}$$

Learning as inference – recap

Maximum posterior (MAP) estimate:

Maximise $\ln p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$

$$\boxed{\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}} \left\{ \frac{\beta}{2} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 + \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \right\}}$$

BUT: Solutions 1 (ML) and 2 (MAP) give point estimates of $\mathbf{w}$

Learning as inference – Bayesian view

Instead of estimating “optimal” parameters $\mathbf{w}$, let’s integrate over all values of $\mathbf{w}$ (let’s make use of the distribution)

Marginalisation:

$$\underbrace{p(t \mid x, \mathbf{x}, \mathbf{t}, \alpha, \beta)}_{\text{predictive}} = \int \underbrace{p(t \mid x, \mathbf{w}, \beta)}_{\text{“noise” model}} \underbrace{p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)}_{\text{posterior}} \, \mathrm{d}\mathbf{w}$$

We assume we know what α and β are.

Learning as inference – Bayesian view

Instead of estimating "optimal" parameters $\mathbf{w}$, let's integrate over all values of $\mathbf{w}$ for our linear model $y = \sum_{i=0}^{M} w_i \phi_i(x)$

For a one-dim polynomial model: $\phi_i(x) = x^i$ (order M-1)

Learning as inference – Bayesian view

Instead of estimating “optimal” parameters $\mathbf{w}$, let’s integrate over all values of $\mathbf{w}$ for our linear model $y = \sum_{i=0}^{M} w_i \phi_i(x)$

For a one-dim polynomial model: $\phi_i(x) = x^i$

$$p(t \mid x, \mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t \mid m(x), s^2(x)\right) \quad \begin{cases} m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \\ s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S}\, \boldsymbol{\phi}(x) \end{cases}$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}}$$

$$\boldsymbol{\phi}(x) = (\phi_0(x), \ldots, \phi_M(x))$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ……..

…..ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \rightarrow \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \rightarrow \mathbf{w}_{\text{MAP}}$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

.....ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \rightarrow \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \rightarrow \mathbf{w}_{\text{MAP}}$$

.....in Bayesian approach $\mathbf{w}$ is integrated over (**marginalisation**)

$$\text{Bayes:} \quad \mathcal{D}, p(\mathbf{w}) \rightarrow p(\mathbf{w} \mid \mathcal{D})$$

c.f. DD2431, Lecture 6, G. Salvi

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

.....ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \to \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \to \mathbf{w}_{\text{MAP}}$$

Use $p(t \mid x_{in}, \mathbf{w}_{\text{ML}})$ or $p(t \mid x_{in}, \mathbf{w}_{\text{MAP}})$ for prediction.

.....in Bayesian approach $\mathbf{w}$ is integrated over (**marginalisation**)

$$\text{Bayes}: \quad \mathcal{D}, p(\mathbf{w}) \to p(\mathbf{w} \mid \mathcal{D})$$

Marginalise over $\mathbf{w}$:

$$p(t \mid \mathcal{D}) = \int p(t \mid x_{in}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, \mathrm{d}\mathbf{w}$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ……..

…..ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \to \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \to \mathbf{w}_{\text{MAP}}$$

Frequentist philosophy

$p(t \mid x_{in}, \mathbf{w}_{\text{MAP}})$ for prediction.

…..in Bayesian approach $\mathbf{w}$ is integrated over (**marginalisation**)

$$\text{Bayes:} \quad \mathcal{D}, p(\mathbf{w}) \to p(\mathbf{w} \mid \mathcal{D})$$

Bayesian philosophy

$$p(t \mid \mathcal{D}) = \int p(t \mid x_{in}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, \mathrm{d}\mathbf{w}$$

Inference and decision (classification)

The ***inference*** stage of classification $\mathcal{D} \rightarrow p(C_k, \mathbf{x}), k = 1,..,K$

Model the inputs $\mathbf{x}$ and outputs C

$$\left.\begin{array}{l} p(\mathbf{x} \mid C_k) \\ p(C_k) \end{array}\right\} \text{ for each class } C_k \qquad p(\mathbf{x}, C_k)$$

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\, p(C_k)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid C_k)\, p(C_k)$$

Inference and decision (classification)

The ***inference*** stage of classification $\mathcal{D} \rightarrow p(C_k, \mathbf{x}), k = 1,..,K$

Remarks:
1) K classes
2) $\mathbf{x}$ – multi-dim input feature vector

Model the inputs $\mathbf{x}$ and outputs C

$$\left.\begin{array}{l} p(\mathbf{x} \mid C_k) \\ p(C_k) \end{array}\right\} \text{for each class } C_k$$

$$p(\mathbf{x}, C_k)$$

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\, p(C_k)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid C_k)\, p(C_k)$$

GENERATIVE approach

Inference and decision (classification)

The ***inference*** stage of classification $\mathcal{D} \to p(C_k, \mathbf{x}), k = 1,..,K$

Model the inputs $\mathbf{x}$ and outputs C

$$\left.\begin{array}{l} p(\mathbf{x} \mid C_k) \\ p(C_k) \end{array}\right\} \text{ for each class } C_k$$

$$p(\mathbf{x}, C_k)$$

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\, p(C_k)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid C_k)\, p(C_k)$$

GENERATIVE approach

Solve first the inference problem of determining posteriors for each class without modelling $p(C_k, \mathbf{x})$ or $p(\mathbf{x} \mid C_k)$

$$p(C_k \mid \mathbf{x})$$

DISCRIMINATIVE approach

Generative vs discriminative approach

What are the virtues of the generative approach?

- The parameters are estimated separately for each class (no need to retrain the model when new classes are added)
- Rather straightforward to fit in a Bayesian framework (but it depends on the problem, sometimes discriminative function can be easier to optimise)

Generative vs discriminative approach

What are the virtues of the generative approach?

- The parameters are estimated separately for each class (no need to retrain the model when new classes are added)
- Rather straightforward to fit in a Bayesian framework (but it depends on the problem, sometimes discriminative function can be easier to optimise)
- Easy and elegant way of handling missing or unlabelled data
- Generative model allows for….. generating data

  -> generative models can be run “backwards”

Generative vs discriminative approach

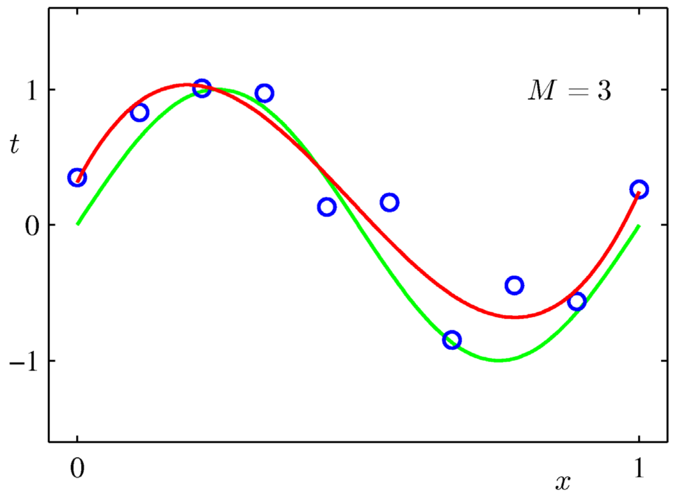
What are the virtues of the generative approach?

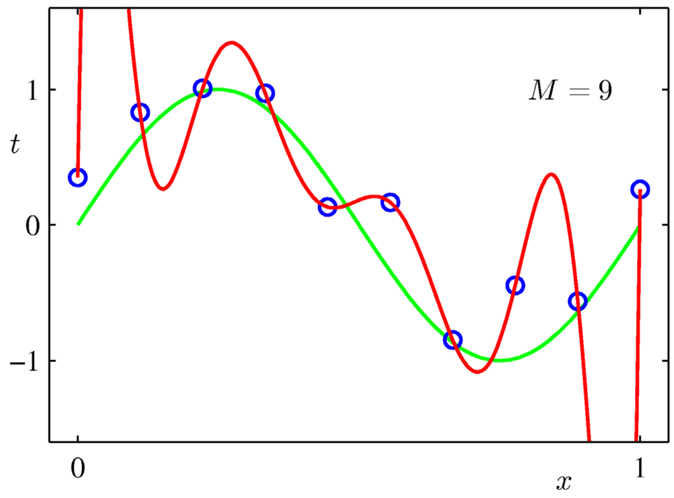
- The parameters are estimated separately for each class (no need to retrain the model when new classes are added)
- Rather straightforward to fit in a Bayesian framework (but it depends on the problem, sometimes discriminative function can be easier to optimise)
- Easy and elegant way of handling missing or unlabelled data
- Generative model allows for….. generating data

  -> generative models can be run “backwards”
- BUT: discriminative models tend to be more accurate (less vulnerable to assumptions)

Model complexity

- Overfitting



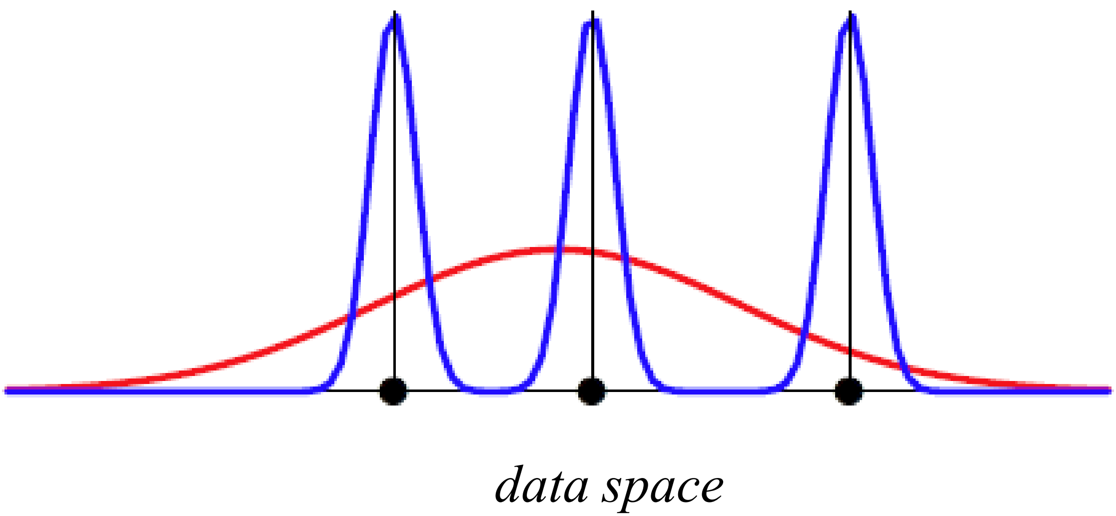


Bishop, sec. 1.1

Please recall from Lecture 1 (J. Lagergren)

Model complexity

- Overfitting



The likelihood can be arbitrarily large by adding parameters.

Please recall from Lecture 6, DD2431 (G. Salvi)

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$

Remarks about notation:

1) $\boldsymbol{x}$ can be a multi-dim input
2) $y(\boldsymbol{x})$ really is $y(\boldsymbol{x},\boldsymbol{w})$ that depends on parameters $\boldsymbol{w}$

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$

- The expected loss, L, to be minimised looks then as follows:

$$\mathbb{E}[L] = \iint \{y(\boldsymbol{x}) - t\}^2 \, p(\boldsymbol{x}, t) \, d\boldsymbol{x} \, dt$$

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$

- The expected loss, L, to be minimised looks then as follows:

$$\mathbb{E}[L] = \iint \{y(\boldsymbol{x}) - t\}^2 \, p(\boldsymbol{x}, t) \, d\boldsymbol{x} \, dt$$

$$\Downarrow$$

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - \mathbb{E}_t[t \mid \boldsymbol{x}]\}^2 \, p(\boldsymbol{x}) \, d\boldsymbol{x} + \int \mathrm{var}[t \mid \boldsymbol{x}] \, p(\boldsymbol{x}) \, d\boldsymbol{x}$$

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$
- The expected loss, L, to be minimised looks then as follows:

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - \mathbb{E}_t[t \mid \boldsymbol{x}]\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} + \underbrace{\int \mathrm{var}[t \mid \boldsymbol{x}]\, p(\boldsymbol{x})\, d\boldsymbol{x}}_{\text{noise on the data, independent of the regressor } y(\boldsymbol{x})}$$

noise on the data,
independent of the regressor $y(\boldsymbol{x})$

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$

- The expected loss, L, to be minimised looks then as follows:

optimal prediction to be accounted for by our regression model $y(\boldsymbol{x})$: $h(\boldsymbol{x}) = \mathbb{E}_t[t \mid \boldsymbol{x}]$

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - \mathbb{E}_t[t \mid \boldsymbol{x}]\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} + \underbrace{\int \mathrm{var}[t \mid \boldsymbol{x}]\, p(\boldsymbol{x})\, d\boldsymbol{x}}$$

$$\{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2$$

noise on the data, independent of the regressor $y(\boldsymbol{x})$

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$

- The expected loss, L, to be minimised looks then as follows:

optimal prediction to be accounted for by our regression model $y(\boldsymbol{x})$: $h(\boldsymbol{x}) = \mathbb{E}_t [t \mid \boldsymbol{x}]$

h is the underlying model (data generating mechanism) that we want to approximate.

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - \mathbb{E}_t[t \mid \boldsymbol{x}]\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} + \underbrace{\int \mathrm{var}[t \mid \boldsymbol{x}]\, p(\boldsymbol{x})\, d\boldsymbol{x}}_{\text{noise on the data, independent of the regressor } y(\boldsymbol{x})}$$

$$\{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2$$

noise on the data,
independent of the regressor $y(\boldsymbol{x})$

Frequentist viewpoint – bias-variance dilemma

- Maximum Likelihood estimate for linear regression corresponds to minimising loss $L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$
- The expected loss, L, to be minimised looks then as follows:

$$\mathbb{E}[L] = \underbrace{\int \{y(\boldsymbol{x}) - \mathbb{E}_t[t \mid \boldsymbol{x}]\}^2 p(\boldsymbol{x})\, d\boldsymbol{x}}_{\text{has to be minimised}} + \underbrace{\int \operatorname{var}[t \mid \boldsymbol{x}]\, p(\boldsymbol{x})\, d\boldsymbol{x}}_{\text{noise on the data, independent of the regressor } y(x)}$$

Frequentist approach

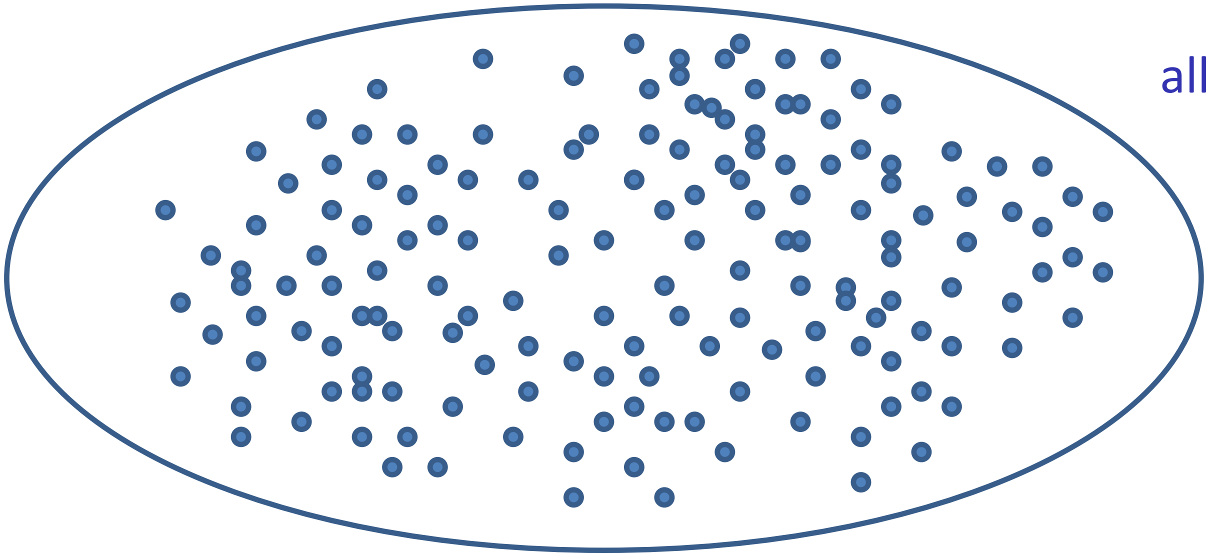
- Find optimal parameters for y using data D

$$\int_{\mathcal{D}} \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} \quad \rightarrow \quad \min \{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2$$

Frequentist approach

- Find optimal parameters for y using data D

$$\int_{\mathcal{D}} \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})\, d\boldsymbol{x} \quad \rightarrow \quad \min \{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2$$

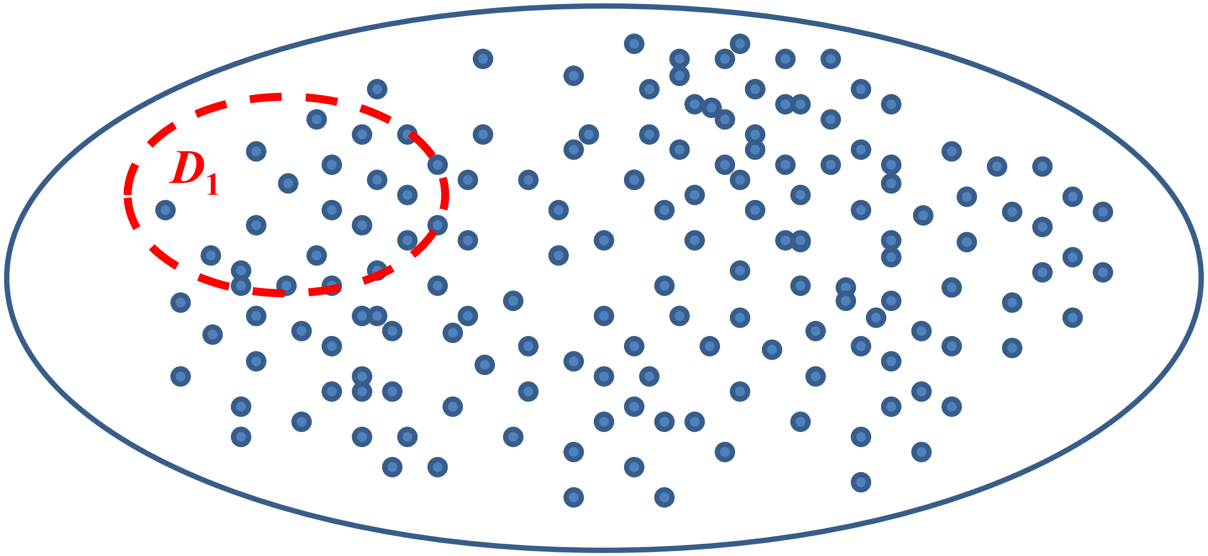


all datasets D

Frequentist approach

- Find optimal parameters for y using data D

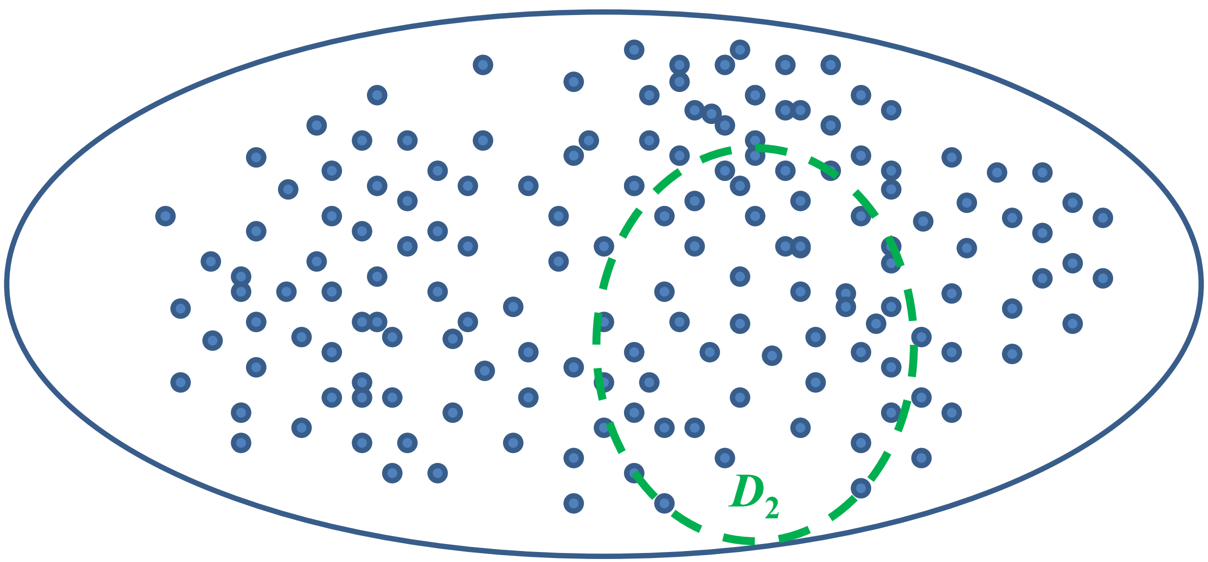
$$\min\left\{y(\boldsymbol{x};\mathcal{D}_1)-h(\boldsymbol{x})\right\}^2$$



Frequentist approach

- Find optimal parameters for y using data D

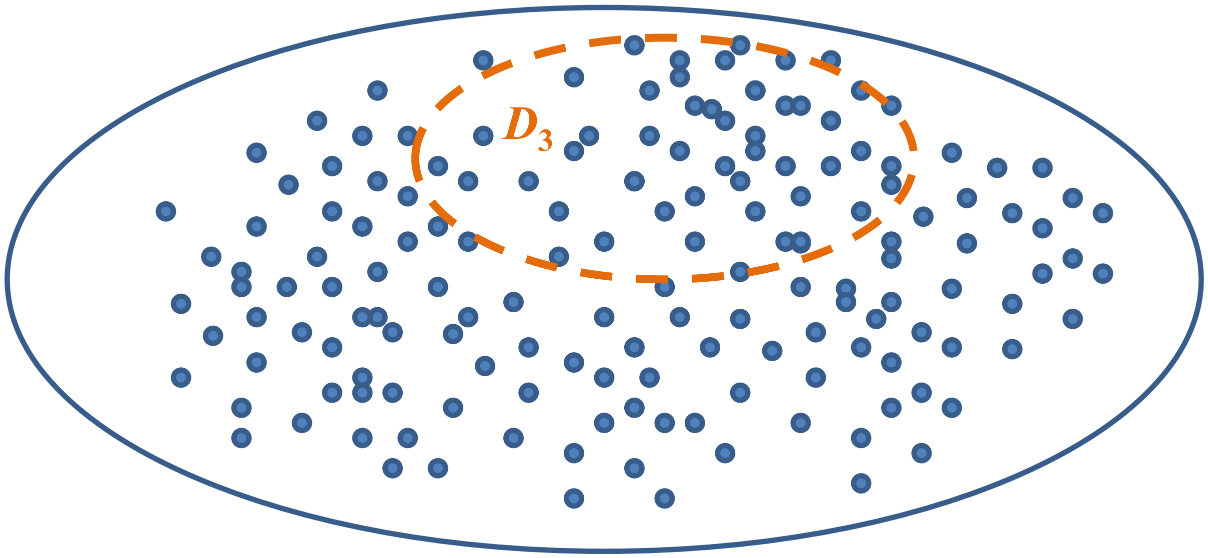
$$\min\left\{y(\boldsymbol{x};\mathcal{D}_2)-h(\boldsymbol{x})\right\}^2$$



Frequentist approach

- Find optimal parameters for y using data D

$$\min\left\{ y(\boldsymbol{x}; D_3) - h(\boldsymbol{x}) \right\}^2$$



Bias-variance decomposition

Expected value over all possible datasets D

$$\mathbb{E}[L] = \mathbb{E}_{\mathcal{D}}\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 + \int \mathrm{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, d\boldsymbol{x}$$

Bias-variance decomposition

Expected value over all possible datasets D

$$\mathbb{E}[L] = \mathbb{E}_{\mathcal{D}}\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 + \int \mathrm{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, d\boldsymbol{x}$$

Taking frequentist perspective of fixing the model for specific data:

$$\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 = \left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] + \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2$$

Bias-variance decomposition

Expected value over all possible datasets D

$$\mathbb{E}[L] = \mathbb{E}_{\mathcal{D}}\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 + \int \mathrm{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, d\boldsymbol{x}$$

Taking frequentist perspective of fixing the model for specific data:

$$\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 = \left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] + \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2$$

…and then averaging all possible data $\mathbb{E}_{\mathcal{D}}$

$$\mathbb{E}_{\mathcal{D}}\left[\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2\right] = \left\{\mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2 + \mathbb{E}_{\mathcal{D}}\left[\left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right]\right\}^2\right]$$

Bias-variance decomposition

Expected value over all possible datasets D

$$\mathbb{E}[L] = \mathbb{E}_{\mathcal{D}}\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 + \int \operatorname{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, d\boldsymbol{x}$$

$$\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 = \left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] + \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2$$

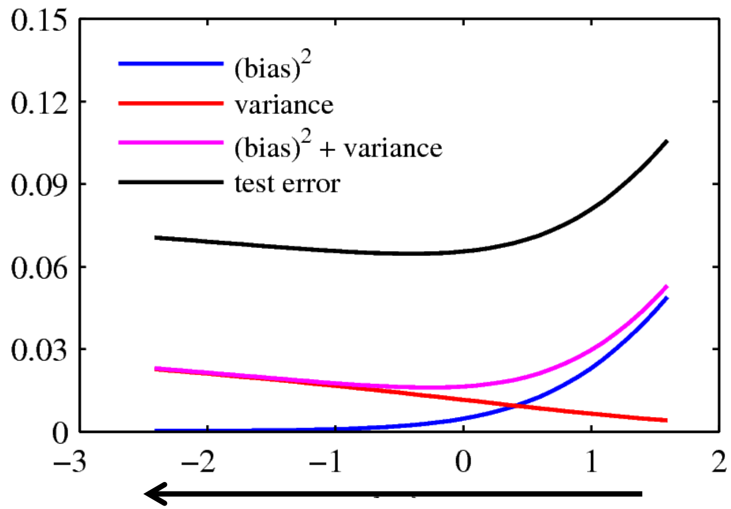
$$\mathbb{E}_{\mathcal{D}}\left[\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2\right] = \underbrace{\left\{\mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right]\right\}^2\right]}_{\text{variance}}$$

Bias-variance dilemma: model complexity

Expected value over all possible datasets D

$$\mathbb{E}_{\mathcal{D}}[L] = \mathbb{E}_{\mathcal{D}}\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 + \int \operatorname{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, dx$$

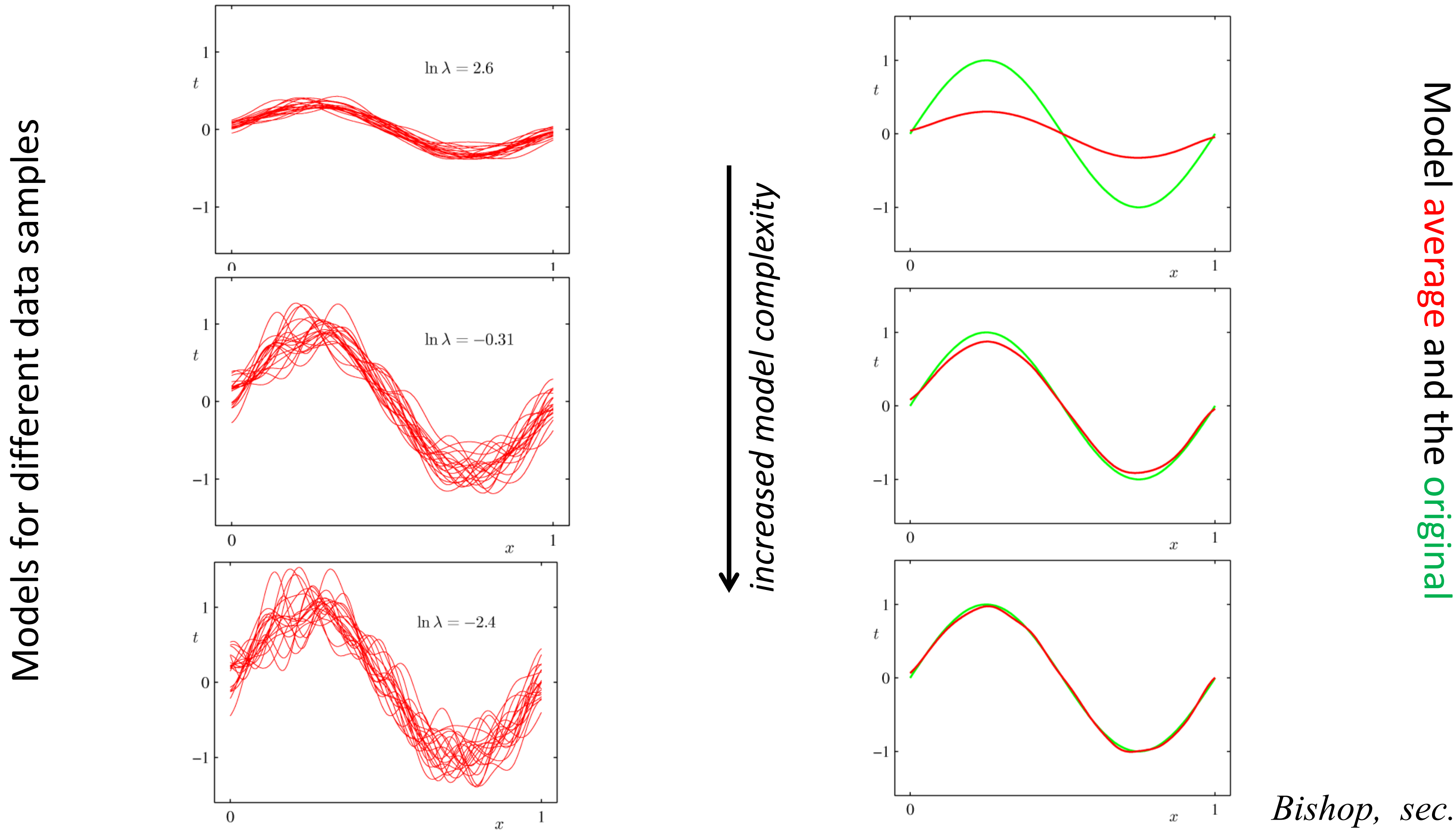
$$\mathbb{E}[L] = (\text{bias})^2 + \text{variance} + \text{noise}$$



increased model complexity

Bishop, sec. 3.2

Bias-variance: illustrative example



Bishop, sec. 3.2

Model selection (*c.f.* L1, J. Lagergren)

- Occam's razor – *"Accept the simplest explanation that fits the data."*

Model selection (*c.f.* L1, J. Lagergren)

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
    - bias-variance dilemma

Model selection (*c.f.* L1, J. Lagergren)

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
  - bias-variance dilemma
  - need to control the model's complexity
    - regularisation
    - correction for the bias of ML estimates (AIC, BIC)
    - empirical estimate of generalisation error on a hold-out set (validation, resampling)
    - structural risk minimization (SRM) (minimise upper bound on the true risk), see also VC dimension (statistical learning theory)

Model selection (*c.f.* L1, J. Lagergren)

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
  - bia
  - ne
    - –
    - –
    - –
    - structural risk minimization (SRM) (minimise upper bound on the true risk), see also VC dimension

Bayesian model selection

(to appear in Lecture 3)

- the second inference stage to compare models the Bayesian way
- Bayes' factor and complexity penalisation (relies on the evidence approximation)

Selected fundamental concepts

1. Introduction – why is probability theory relevant in ML?
2. Probabilistic approach to ML – principles (*recap*).
3. **Probability basics.**

Selected fundamental concepts

1. Introduction – why is probability theory relevant in ML?
2. Probabilistic approach to ML – principles (*recap*).
3. **Probability basics.**

You should really be familiar with the fundamentals!

Please take a closer look at sec. 1.2, 2.1-2.4 in (Bishop, 2006)

Here, we will VERY briefly mention

- the multivariate Gaussian distribution
- Central Limit Theorem

Selected fundamental concepts

1. Introduction – why is probability theory relevant in ML?
2. Probabilistic approach to ML – principles (*recap*).
3. **Probability basics.**

<u>You should really be familiar with the fundamentals!</u>

Plea

Her

More to come in
the first practical session

focus on sec.2.3- 2.4 (Bishop)

Central Limit Theorem

The Gaussian (normal) distribution

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{L/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

$\mathbf{x} \in \mathbb{R}^L$

$\boldsymbol{\mu}$ – mean vector, $\mathbb{R}^L$

$\boldsymbol{\Sigma}$ – covariance matrix , L x L

The Gaussian (normal) distribution

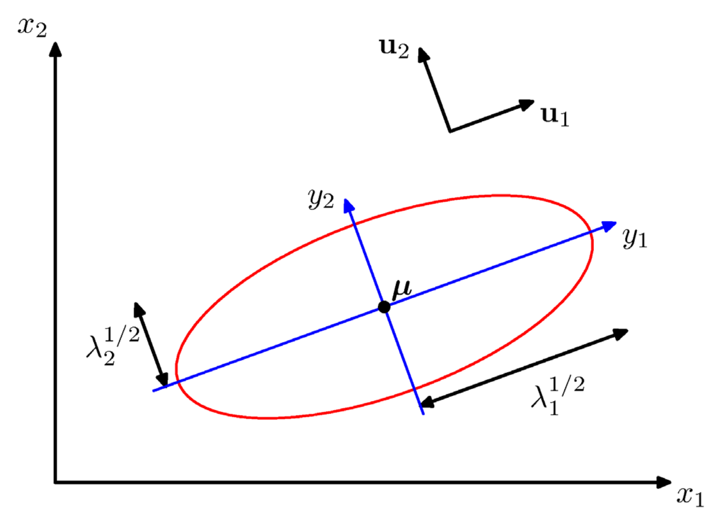
$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{L/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \overbrace{\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}}^{\text{quadratic form}}$$

$\mathbf{x} \in \mathbb{R}^L$

$\boldsymbol{\mu}$ – mean vector, $\mathbb{R}^L$

$\boldsymbol{\Sigma}$ – covariance matrix , L x L

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$



Bishop, sec. 2.3

The Gaussian (normal) distribution

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{L/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\overbrace{\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}}^{\text{quadratic form}}$$

$\mathbf{x} \in \mathbb{R}^L$

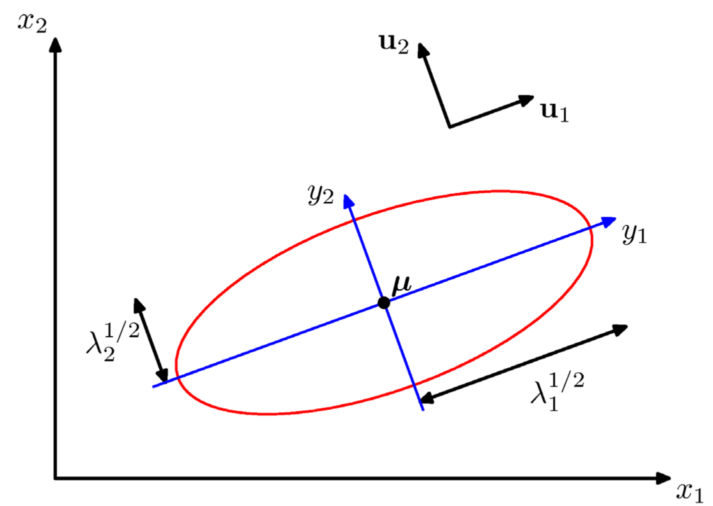
$\boldsymbol{\mu}$ – mean vector, $\mathbb{R}^L$

$\boldsymbol{\Sigma}$ – covariance matrix , L x L

$$\mathbb{E}[\mathbf{x}] = \int \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathbf{x}\, d\mathbf{x} = \boldsymbol{\mu}$$

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$



Bishop, sec. 2.3

The Gaussian distribution

Two-dimensional (2D) case, i.e. $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

<u>Remark about notation</u>:

Unlike in the beginning, where x_i corresponded to the i-th realisation of one-dim r.v. x, **here** x_1 and x_2 correspond to one-dimensional variables that constitute two-dim r.v. $\mathbf{x}$.

The Gaussian distribution

Two-dimensional (2D) case, i.e. $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\mu_i = \mathbb{E}[x_i]$$

$$\sigma_{i,j} = \sigma(x_i, x_j) = \mathbb{E}\left[\left(x_i - \mathbb{E}[x_i]\right)\left(x_j - \mathbb{E}[x_j]\right)^{\mathrm{T}}\right]$$

$$\sigma_{i,i} = \sigma_i^2 = \mathrm{var}[x_i]$$

The Gaussian distribution

Two-dimensional (2D) case, i.e. $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

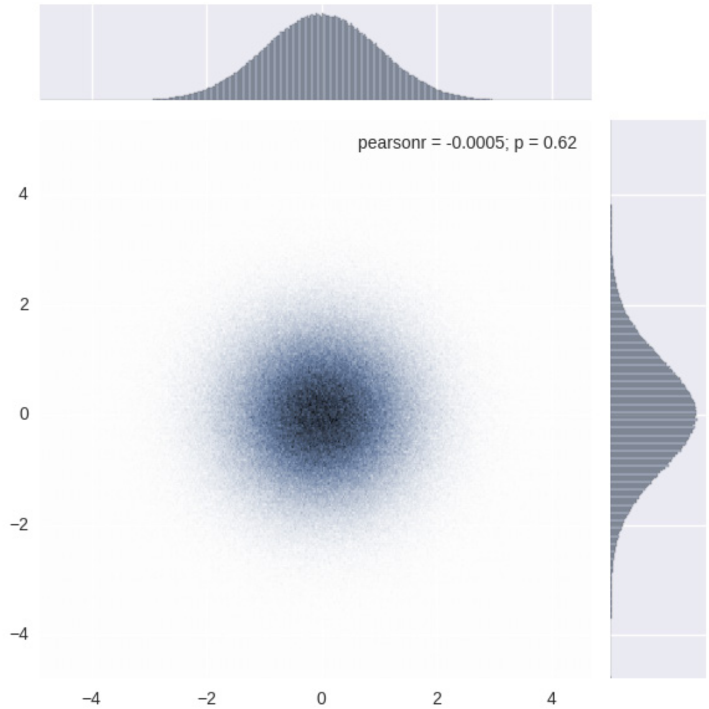
$$\mu_i = \mathbb{E}[x_i]$$

$$\sigma_{i,j} = \sigma(x_i, x_j) = \mathbb{E}\left[\left(x_i - \mathbb{E}[x_i]\right)\left(x_j - \mathbb{E}[x_j]\right)^{\mathrm{T}}\right]$$

$$\sigma_{i,i} = \sigma_i^2 = \mathrm{var}[x_i]$$

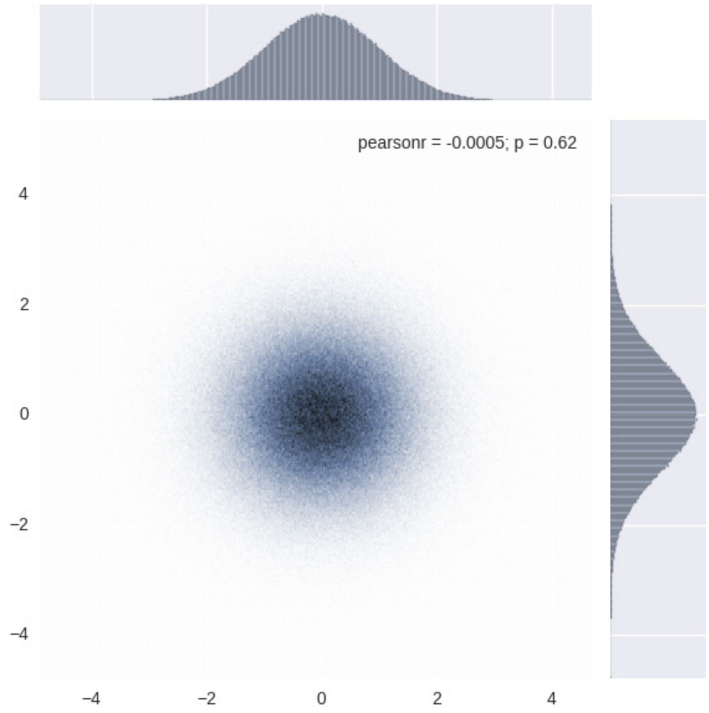
Sample covariance: $\bar{\sigma}_{i,j} = \dfrac{1}{N-1} \sum_{k=1}^{N} \left(x_i^{(k)} - \mu_i\right)\left(x_j^{(k)} - \mu_j\right)$

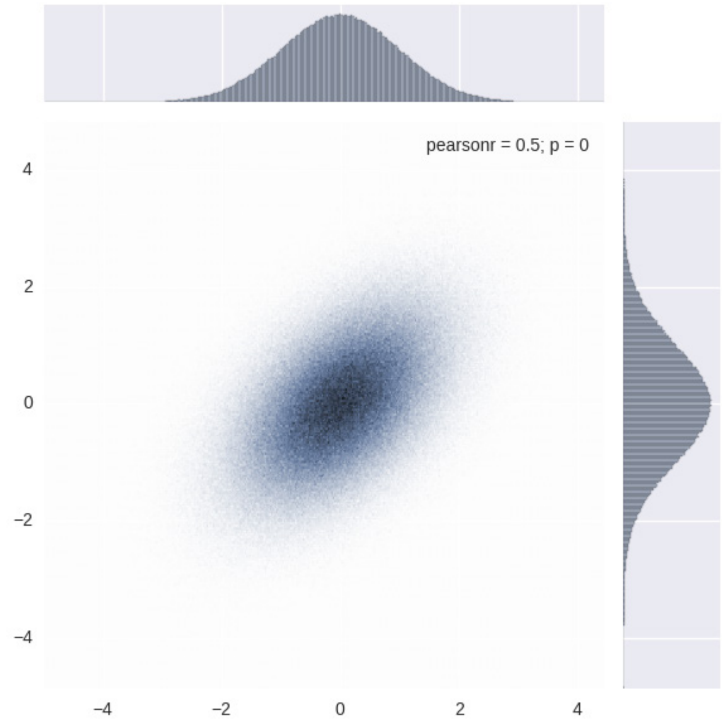
Covariance – 2D case



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Covariance – 2D case





Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Central Limit Theorem

"The sum of a number of independent and identically distributed random variables with finite variances will tend to a normal distribution as the number of variables grows."

$$\mu = \mathbb{E}\left[X_i\right]$$
$$\sigma^2 = \operatorname{var}\left[X_i\right]$$

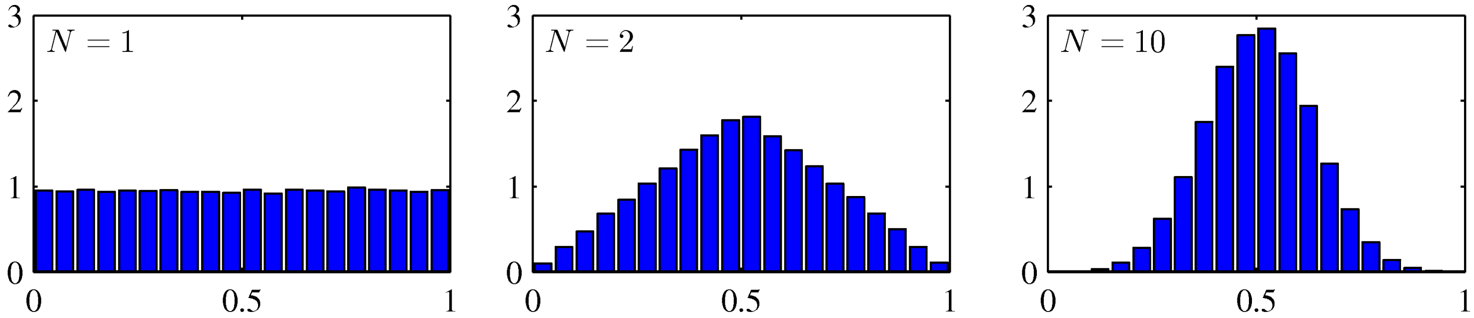
$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathcal{N}\left(\mu, \sigma/\sqrt{n}\right)$$

Central Limit Theorem

"The sum of a number of independent and identically distributed random variables with finite variances will tend to a normal distribution as the number of variables grows."

$$\mu = \mathbb{E}[X_i]$$
$$\sigma^2 = \mathrm{var}[X_i]$$

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n \to \infty} \mathcal{N}\left(\mu, \sigma/\sqrt{n}\right)$$



Bishop, sec. 2.3